DOCUMENT RESUME

ED 159 222 TM 007 704

AUTHOR Smith, Douglas U.
TITLE The Effects of Various Item Selection Methods on the Classification Accuracy and Classification Consistency of Criterion-Referenced Instruments.
PUB DATE Mar 78
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; *Item Analysis; Item Sampling; Mastery Tests; *Mathematical Models; *Test Construction; *Test Reliability; *Test Validity
IDENTIFIERS Latent Trait Theory

ABSTRACT

This study examined the effects of certain item selection methods on the classification accuracy and classification consistency of criterion-referenced instruments. Three item response data sets, representing varying situations of instructional effectiveness, were simulated. Five methods of item selection were then applied to each data set for the purpose of selecting a best subset of items. Parallel forms of an instrument were then constructed and two measures of test quality obtained: the proportion of correct classifications, and the mean split-half coefficient of agreement. Results indicated that methods yielding the best accuracy and consistency varied across the situations of instructional effectiveness. When a "fair amount" of variability is present in the distribution of test scores, as was the case in Situation 3, the point-biserial correlation method was superior to the other methods. However, in true mastery learning situations, approximated in Situation 2, the Cox and Vargas method of item selection is to be recommended. Finally, in those situations where instruction is relatively ineffective, as was the case with Situation 1, the Brennan and the phi coefficient methods tended to yield the best instruments. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The Effects of Various Item Selection Methods on the Classification Accuracy and Classification Consistency of Criterion-Referenced Instruments

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Douglas Smith

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Douglas U. Smith

Division of Educational Services & Research

Wayne State University School of Medicine

Detroit, Michigan

ED159222

Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada, March 1978.

Printed in U.S.A.

TM007 704

Abstract

This study examined the effects of certain item selection methods on the classification accuracy and classification consistency of criterion-referenced instruments. Three item response data sets, representing varying situations of instructional effectiveness, were simulated. Five methods of item selection were then applied to each data set for the purpose of selecting a best subset of items. Parallel forms of an instrument were then constructed and two measures of test quality obtained: the proportion of correct classifications, and the mean split-half coefficient of agreement. Results indicated that methods yielding the best accuracy and consistency varied across the situations of instructional effectiveness.

The Effects of Various Item Selection Methods on the Classification Accuracy and Classification Consistency of Criterion-Referenced Instruments

Introduction

When assessment instruments are used as tools to aid in educational decision-making, concern over the adequacy of the instrument should focus on two issues. The first involves the extent to which the instrument is capable of assigning individuals to their true level of performance (e.g. pass or fail, mastery or nonmastery, acceptable or unacceptable performance). A second concern involves the degree to which these individual decisions are consistent across repeated administrations of the instrument. Clearly, an instrument which classifies an examinee correctly on one occasion, but fails to do so on another is as inadequate for decision-making as one which consistently misclassifies examinees.

Typically, such forms of assessment are often referred to as being criterion-referenced (CR). As a result of such assessments, individuals may be classified into, say, a mastery category if their performance equals, or exceeds some predetermined level of minimum performance, otherwise the performance may be taken as evidence of nonmastery. It is this type of assessment which will be the central focus of this paper.

A major goal of good CR measurement would seem to be aimed at maximizing the proportion of correct mastery-type decisions which are made regarding a group of examinees, while, at the same time, providing a high degree of classification consistency (across repeated administrations of the instrument).

In order for this to be accomplished consideration must be given to the quality of those items which are selected from an available pool in order to insure that the items comprising the final form of the assessment instrument are able to effectively discriminate, say, between true masters and nonmasters. This latter concern however, must be subject to the restriction that the selected set of items is representative of the larger domain of which it was intended to measure. Thus, a major issue confronting the developer of CR instruments becomes one of selecting items, which, when taken collectively, constitute instruments capable of yielding both accurate as well as consistent judgments regarding an individual's true level of mastery, while at the same time being representative of the domain of interest.

The use of item selection methods for CR test construction has come under some criticism. A number of investigators, notably Cox (1965), Millman and Popham (1974), and Hambleton, Swaminathan, Algina and Coulson (1975) have argued against the application of empirical methods of item selection on the grounds that such methods may destroy the all-important content validity of these tests.

Yet, a number of points may be made in the defense of these procedures. The first is that it seems presumptous, at least to this investigator, that each item comprising a domain of items measures it equally as well. Most certainly, items which have been written to assess an individual's performance with respect to some domain will vary not only in terms of difficulty, but also to the extent which they relate to the domain of interest. A second point is that the use of item selection procedures does not necessarily have to affect the content validity of an instrument. For example, a test developer could select only the most highly discriminating items from each of the categories comprising the original test plan, while managing to retain those items from each category in the same proportions as exist with-

in the original item pool. As a final point, Wedman (1973) has noted that empirical methods of item selection may, in fact, enhance test characteristics by helping to alleviate some of the subjective judgments which may enter into the item writing process.

Another issue relevant to the selection of items for inclusion into a final form of a CR instrument involves the actual identification of "bad" items. For example, when an item analysis is performed on a set of items administered following instruction, it is difficult, if not near impossible, to assess whether it is the item, or the instruction, which is faculty.

Yet, it should be noted that any evaluation of the adequacy of item selection methods for CR test development must employ an appropriate criterion. Since a major purpose of CR measurement involves the classification of individuals into one of several mutually exclusive categories (Crehan, 1974; Millman & Popham, 1974; Swaminathan, Hambleton & Algina, 1975), a reconceptualization of test score reliability and validity appears to be in order. This investigator concurs with Hambleton and Novick (1973) who view CR test score validity from the standpoint of the accuracy by which individual mastery decisions are made, and reliability in the sense of the replicability of the decisions (Carver, 1970). In short, it must be these decisions which are both valid (i.e. accurate) and reliable (i.e. consistent).

The present investigation, therefore, attempted to answer the following questions, each of which was examined under varying degrees of instructional effectiveness. They were: (1) How do instruments composed of items identified as being "best" according to a particular method of item selection differ with respect to classifying examinees into their true mastery state? (2) How consistent are such instruments in their classification of individuals?

Procedure

Item Response Simulation

The responses of a group of 1000 "examinees" to a set of 70 items administered both as a pretest and again as a posttest under varying degrees of instructional effectiveness were simulated in accordance with a two-parameter latent trait model. A computer program written expressly for this type of simulation was employed. Details of this computer program may be found in Hambleton and Rovinelli (1973).

With this model items are fully described by means of two parameters: item difficulty and item discrimination. For use in the present simulation values of the item difficulty parameter were sampled from a unifrom distribution having a lower bound of -3 and an upper bound of +3. Values of the item discrimination parameter, on the other hand, were sampled from a normal distribution having a mean of 1.0 and a standard deviation of .75. Since these parameter values remain relatively invariant despite systematic changes occurring within a group of examinees, item characteristics were held constant from pre to posttesting, while examinee ability was varied to reflect varying degrees of instructional effectiveness.

Inital (pretest) specification of examinee ability was made by sampling values from a normal distribution having a mean of -1.5 and a standard deviation of 1.0. Additionally, the distribution of pretest ability was restricted in the sense that θ_n was confined to the interval -3 to +3. Having specified the values for both of the item parameters as well as examinee ability at the time of pretesting, pretest examinee item responses were generated by means of the aforementioned computer program.

In simulating the responses of the group of examinees to the same set of items administered following instruction, a linear gain model was employed. This model was used to relate examinee ability following instruction to hir ability at the time of pretesting by means of the expression: $\theta_{n,\,post} = ka_n + \theta_{n,\,pre}$, where k was a constant for each examinee; a_n, a value unique to each examinee, and $\theta_{n,\,pre}$, the latent ability of the examinee at the time of pretesting.

Specifically, three forms of this model were used in simulating varying degrees of instructional effectiveness. These variations, differing with respect to the mean amount of change in examinee ability (group) from pretesting to posttesting were used to simulate varying degrees of instructional effectiveness. These situations were designed to represent the following: relatively little change in examinee ability between the time of pretesting and posttesting (ineffective instruction), substantial gain in examinee ability from pretesting to posttesting (effective instruction), and a third situation in which examinees who were initially higher in ability on the pretest (based on a median split) gained more as a result of instruction than persons of lower initial (pretest) ability. Each of the resultant data sets, hereafter referred to as Situations 1 - 3 respectively, consisted of the responses of 1000 "examinees" to the set of 70 items.

Item Selection Methods

Each data set was then subjected to an item analysis. Values of four item selection indices were computed. They were: (1) the point-biserial correlation (RPB) between the dichotomous item response (pass, fail) and the total test score, (2) the difference between the posttest and pretest proportion of examinees answering the item correctly (CV) (Cox & Vargas, 1966), (3) the difference between the proportion of examinees in the mastery group

And the nonmastery group passing the item (BR) (Brennan, 1972), and (4) the phi coefficeint (PHI) computed between the dichotomous item response versus mastery/nonmastery status on the total test. A fifth method of item selection, based on the random selection of items (RDM), was also employed.

Instrument Construction

The results of the item analysis of the three data sets were used to identify a "best" subset of 40 items from the original item pool. From each of these sets of "best" items, ten randomly parallel forms of a CR assessment instrument were constructed. Each test form consisted of a randomly selected subset of 20 items drawn from the group of 40 items and, as a result, became the unit of analysis for assessing the classification accuracy and classification consistency of CR instruments.

Classification Accuracy

For each instrument constructed classification accuracy was assessed by means of the proportion of examinees who were correctly classified into their true mastery state. True mastery was determined for each examinee by computing their true score (TS) on the original set (i.e. domain) of 70 items, using the expression: $TS = (1 / k) \sum_{g=1}^{k} P_g (0_n)$, where k is equal to the number of items (70), and $P_g (0_n)$ represents the probability of success of examinee n on item g. Thus, by comparing an examinee's TS to the level of minimum mastery (MPL), which was set at 80 per cent correct in the present study, an examinee's true mastery status could be determined. The following classification rule was employed: if an examinee's TS was greater than or equal to the MPL, then true mastery status was granted; otherwise it was denied.

A similar assessment of individual mastery was made on the basis of individual performance on the "constructed tests". Again, using a 80 per cent

mastery level. Thus, for each examinee in Situations 1 - 3 both a true mastery level as well as an estimate of their mastery status based on performance on the constructed instruments was obtained. The proportion of examinees who were correctly classified into their true mastery state by each instrument was used as the measure of the classification accuracy of the instrument.

Classification Consistency

In a similar manner a measure of each instrument's ability to consistently classify examinees was obtained. The particular measure of classification consistency used in this investigation was that of the mean split-half coefficient of agreement, coefficient beta originally described by Marshall and Haertel (1975). Although there exist other measures which can be used to assess the decision-making consistency of CR instruments (Carver, 1970; Swaminathan, Hambleton & Algina, 1975; Subkoviak, 1976; and Huynh, 1976), coefficient beta was favored for two reasons. First, it may be determined from only a single administration of an instrument, and secondly, it is conceptually similar to the more traditional KR-20 coefficient of internal consistency reliability.

Results

The research design employed in this investigation is displayed in Figure 1. The independent variable was the type of item selection method employed, while the dependent variables of interest were classification accuracy, as measured by the proportion of correct classifications, and classification consistency as measured by the mean split-half coefficient of agreement (coefficient beta). Separate analyses were carried out for each of the three situations of instructional effectiveness. The unit of analysis was

a "form" of a CR assessment instrument resulting from the random selection of 20 items from a set of 40 "best" items as identified by a particular selection method. For each selection method 10 such "forms" (replications) were generated.

Insert Figure 1 about here

A Kruskal-Wallis one-way analysis of variance of ranked data was performed on the proportion of correct classifications for each of the five methods of item selection. The results of this analysis appear in Table 1. Seperate analyses were carried out for each of the three situations of instructional effectiveness. Values of the rank sums for the proportion of correct classifications (across 10 replications) is given along with the value of the Kruskal-Wallis H statistic (corrected for ties). It should be noted that a large value for the rank sum indicates a higher degree of classification accuracy. It can be observed in this table that the obtained value of the test statistic exceeds the critical value (α = .05) for each of the three situations. Thus, the null hypothesis of no difference between the mean rankings of the five methods of item selection can be rejected.

Insert Table 1 about here

Pairwise comparisons were then performed on the mean rankings for each of the five selection methods. The procedure employed in this analysis was one attributable to Dunn (1964). This method allowed for the simultaneous comparison of any pair of mean ranks, while, at the same time, keeping the "situation wise" error rate constant (α = .05). The results of this analysis

are presented in Table 2. It should be noted in this table that the method of item selection which appears on the extreme left is the one which yielded the highest degree of classification accuracy.

Insert Table 2 about here

For Situation 1 (ineffective instruction) the CV method yielded the poorest classification accuracy, and was significantly poorer than any of the other four methods. The analysis for Situation 2 (effective instruction) reveals three homogeneous subsets. Closer inspection of these data indicates that instruments constructed by the CV method were significantly better in terms of classification accuracy than instruments developed in accordance with other methods. In Situation 3 (differentially effective instruction) the best classification accuracy was obtained via the RPB method. Although not statistically better than the CV method, the aforementioned methods were found to be better in terms of their classification accuracy than either the BR or PHI methods.

A similar approach to the analysis of classification consistency was taken. Table 3 displays the results of this analysis in which the mean split-half coefficient of agreement was used as the dependent measure. As may be noted in this table, statistically significant differences between the three situations of instructional effectiveness were observed. Again, as indicated in this table, a high value of the rank sum indicates a higher degree of classification consistency.

Insert Table 3 about here

As was the case in the preceeding analysis (classification accuracy), an attempt was made to locate those pairwise differences in the rank sums which contributed to the overall significance of the test statistic. The Dunn method was, once again, employed. Table 4 presents the results of this analysis.

Insert Table 4 about here

The analysis of Situation 1 revealed that the PHI method of item selection produced instruments whose classification consistency was significantly better than either the RPB, CV, or RDM methods. Likewise the BR method yielded a higher degree of classification consistency than did either the CV or RDM methods. The remaining comparisons were, in a statistical sense, nonsignificant.

Finally, Table 5 attempts to summarize the results of the analysis of classification accuracy and classification consistency. The entries in this table correspond to the overall rankings of the rank sums for each method of item selection under the three situations of instructional effectiveness.

Insert Table 5 about here

Discussion

Results of the analysis of classification accuracy revealed that no single method of item selection resulted in instruments which were consistently superior to other methods in correctly classifying examinees when

instructional effectiveness was varied. Specific differences in the selection methods have already been noted.

A similar finding was observed with respect to the measure of classification consistency. That is, no particular method of item selection was consistently superior to others under the three situations of instructional effectiveness. Inspection of Table 5, furthermore, reveals that with the exception of Situation 3 (differentially effective instruction), similar rankings with respect to the measures of interest were not found.

Overall, the findings reported in this paper suggest that the classification accuracy and classification consistency of CR assessment instruments can be improved through the use of item selection methods. When a "fair amount" of variability is present in the distribution of test scores, as was the case in Situation 3, the RPB method was superior to the other methods. However, in true mastery-learning situations, approximate by Situation 2, the CV method of item selection is to be recommended. Finally, in those situations where instruction is relatively ineffective, as was the case with Situation 1, the BR and PHI methods tended to yield the best instruments.

References

Brennan, R.L. A generalized upper-lower item discrimination index. Educational and Psychological Measurement, 1972, 32, 289-303.

Carver, R.P. Special problems in measuring change with psychometric devices. In Evaluative research: Strategies and methods. Washington, D.C.: American Institutes for Research, 1970.

Cox, R.C. Item selection techniques and evaluation of instructional objectives. Journal of Educational Measurement, 1965, 2, 181-185.

Cox, R.C., & Vargas, J.S. A comparison of item-selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1966. (ERIC Document Reproduction Service No. ED 010 517)

Crehan, J.D. Item analysis for teacher-made mastery tests. Journal of Educational Measurement, 1974, 11, 255-262.

Dunn, O.J. Multiple comparisions using rank sums. Technometrics, 1964, 6, 241-252.

Hambleton, R.K., & Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hambleton, R.K., & Rovinelli, R. A FORTRAN IV program for generating examinee response data from logistic test models. Behavioral Science, 1973, 18, 74.

Hambleton, R.K., Swaminathan, H., Algina, J., & Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. An invited symposium presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1975. (ERIC Document Reproduction Service No. ED 107 722)

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Marshall, J.L., & Haertel, E.H. A single-administration reliability index for criterion-referenced tests: The mean split-half coefficient of agreement. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1975. (ERIC Document Reproduction Service No. ED 118 618)

Millman, J., & Popham, W.J. The issue of item and test variance for criterion-referenced tests: A clarification. Journal of Educational Measurement, 1974, 11, 137-138.

Subkoviak, M. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976 13, 265-276.

Swaminathan, H., Hambleton, R.K., & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. Journal of Educational Measurement, 1975, 12, 97-98.

Wedman, I. Reliability, validity, and discrimination measures for criterion-referenced tests. Educational Reports, Umea, No. 4, 1973.

Figure 1. Research Design

Situation (1 to 3)				
Selection Method				
RPB	CV	BR	PHI	RDM
R_1	R_1	R_1	R_1	R_1
R_2	R_2	R_2	R_2	R_2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
R_{10}	R_{10}	R_{10}	R_{10}	R_{10}

Table 1

Kruskal-Wallis Analysis of the Five Methods

of Item Selection for Classification Accuracy

	Rank Sum for Each Selection Method					
Situation	RPB	CV	BR	PHI	RDM	H^a
1	257.50	71.00	375.50	257.50	313.50	24.41*
2	185.50	449.00	151.00	151.00	338.50	33.45*
3	419.50	371.00	122.50	92.50	269.50	39.87*

[a]Corrected for ties

*$p < .05$; χ^2 (4; .95) = 9.488

Table 2

Results of the Multiple Comparisons Between the Mean Ranks for Classification Accuracy

	Selection Method				
	BR	RDM	RPB	PHI	CV
Situation 1	37.55	31.35	25.75	25.75	7.1
	CV	RDM	RPB	BR	PHI
Situation 2	44.9	33.85	18.55	15.1	15.1
	RPB	CV	RDM	BR	PHI
Situation 3	41.95	37.1	26.95	12.25	9.25

Note. Means ranks underscored by the same line are homogeneous.

Table 3

Kruskal-Wallis Analysis of the Five Methods

of Item Selections for Classification Consistency

	Rank Sum for Each Selection Method					
Situation	RPB	CV	BR	PHI	RDM	H^a
1	207.00	145.00	361.50	436.50	125.00	35.58*
2	214.50	447.00	235.50	204.50	173.50	22.59*
3	428.00	365.00	152.00	113.00	216.50	34.89*

[a]Corrected for ties

*$p < .05$; χ^2 (4; .95) = 9.488

Table 4

Results of the Multiple Comparisons Between

the Mean Ranks for Classifications Consistency

	Selection Method				
	PHI	BR	RPB	CV	RDM
Situation 1	43.65	36.15	20.7	14.5	12.5
	CV	BR	RPB	PHI	RDM
Situation 2	44.7	23.55	21.45	20.45	17.35
	RPB	CV	RDM	BR	PHI
Situation 3	42.8	36.5	21.65	15.2	11.3

Note. Means ranks underscored by the same line as homogeneous.

Table 5

Overall Rankings of Item Selection Methods

for Classification Accuracy and Consistency

Situation	Dependent Measure	Selection Method				
		RPB	CV	BR	PHI	RDM
1	Classification Accuracy	3.5	5	1	3.5	2
	Classification Consistency	3	4	2	1	5
2	Classification Accuracy	3	1	4.5	4.5	2
	Classification Consistency	3	1	2	4	5
3	Classification Accuracy	1	2	4	5	3
	Classification Consistency	1	2	4	5	3